European Journal of Sustainable Development Research
2026, 10(3), em0397
e-ISSN: 2542-4742
https://www.ejosdr.com/

MODESTUM

Review Article

OPEN ACCESS

Advancing sustainable development goals through multilingual text summarization: A transformer-based approach

Atul Kumar [1,2*], Shashi Kant Gupta [1,3], Ratan Singh Yadav [4], Uma Shankar Yadav [5], Shekhar Saroj [5], Vivek Kumar [6]

[1] Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, MALAYSIA
[2] School of Computer Science Engineering, Chandigarh University Uttar Pradesh, Unnao-209859, INDIA
[3] Center for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, INDIA
[4] MNNIT-Allahabad, Prayagraj, INDIA
[5] Postdoctoral Researcher, Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, MALAYSIA
[6] Computer Science and Engineering, Bennett University, Greater Noida, INDIA
***Corresponding Author:** pdf.atulkumarverma@lincoln.edu.my

Citation: Kumar, A., Gupta, S. K., Yadav, R. S., Yadav, U. S., Saroj, S., & Kumar, V. (2026). Advancing sustainable development goals through multilingual text summarization: A transformer-based approach. *European Journal of Sustainable Development Research, 10*(3), em0397. https://doi.org/10.29333/ejosdr/18328

ARTICLE INFO

Received: 08 Sep. 2025

Accepted: 18 Feb. 2026

ABSTRACT

This study uses a multilingual summarization model that is based on transformers to justify the United Nations (UN) sustainable development goals (SDGs). The multilingual and immense size of the information related to sustainability renders the synthesis of the related insights challenging policymakers, researchers, and NGOs. To handle this difficulty, we reconfigured a multilingual pre-trained transformer model (mT5) on a new customized set of sustainability texts and news articles, UN reports, and NGO reports in English, Spanish, and French. The methodology flow included the dataset generation, preprocessing, fine-tuning the model, and its thorough evaluation with the help of the automated (ROUGE) and human evaluation. The fine-tuned mT5 was found to be more semantically covered and contextually relevant than the extractive baseline that had the highest ROUGE-L and ROUGE-2 by quantitative analysis, indicating that it achieved a 42% and 39% higher ROUGE-L and ROUGE-2, respectively. The summary generated using human evaluators was rated with a higher level of coherence, fluency, factual accuracy, and SDG relevance (average score 4.2/5) as compared to extractive methods (2.8/5). The outcomes steady the theory that the fine-tuning of domain-specific multilingual processes yields significant improvement in the quality and reliability of automated sustainability summaries. The suggested framework exemplifies the way multilingual natural language processing (NLP) can facilitate access to cross-lingual knowledge and speed up the process of evidence-based decision-making in favor of the SDGs.

Keywords: abstractive summarization, multilingual NLP, natural language processing, sustainable development goals, text summarization, transformer models

INTRODUCTION

The United Nations' (UN) 2030 agenda for sustainable development establishes 17 sustainable development goals (SDGs) aimed at addressing global challenges such as poverty, climate change, environmental degradation, gender inequality, and unequal access to education (UNDP, 2015). Achieving these goals requires timely and accessible information that can support evidence-based decision-making. However, sustainability documents, produced by governments, international organizations, NGOs, and research institutions, are often long, technical, and written in multiple languages. This makes it difficult for stakeholders to efficiently understand and compare information across regions and linguistic contexts (Luo et al., 2022; van der Ploeg & Osei, 2023; Webersinke et al., 2022).

Automatic text summarization (ATS), a core task in natural language processing (NLP), offers a potential solution by condensing lengthy documents into concise summaries while preserving essential meaning (Erkan & Radev, 2004; Luhn, 1958). Early ATS methods mainly relied on extractive techniques that selected key sentences from the source text (Bahdanau et al., 2015; Sutskever et al., 2014; Vaswani et al., 2017). With advances in deep learning and transformer-based architectures (Lewis et al., 2020; Liu & Lapata, 2019; Raffel et al., 2020; Zhang et al., 2020) abstractive summarization has become possible, enabling the generation of new sentences that capture the semantic content of the original document.

Despite these advancements, current summarization models face key limitations. Most state-of-the-art systems are trained on general-purpose, English-only datasets and are not adapted to the specialized vocabulary, dense factual content, and policy-focused structure of sustainability documents. Furthermore, sustainability information is inherently multilingual (Bhat et al., 2023; Ladhak et al., 2020), yet existing multilingual models often struggle with cross-lingual generalization and domain-specific terminology. Popular multilingual summarization datasets also lack SDG-aligned content, creating a gap between real-world needs and available resources.

Multilingual transformer models such as mT5 (Xue et al., 2021), provide an opportunity to address these gaps, but they require domain-specific adaptation to perform effectively on SDG-related texts. To meet this need, the present study introduces a multilingual abstractive summarization framework fine-tuned on a newly curated dataset, SDG-Summ, which includes sustainability documents in English, Spanish, and French. The goal is to generate summaries that are coherent, factually accurate, and relevant to SDG themes across languages.

Research Gap and Theoretical Motivation

Although ATS and multilingual NLP are improving fast, current studies demonstrate that there are still a number of open theoretical and methodological gaps when the methods are utilized on SDG-oriented documents.

First, the majority of the summarization models are trained and tested in domain-neutral conditions, most often with English news collection materials including CNN/DailyMail and XSum. These datasets are very narrative and ossuaries, but do not correspond to the policy-based, indicator-laden, and fact-filled format of sustainability reports. Theoretically, this causes a mismatch of the domains between training goals and the actual SDG communication requirements in the real world where the summaries should avoid losing policy intent, quantitative indicators, and cross-goal relationships. Current models hence do not have the representational basis needed in sustainability specific summarization.

Second, multilingual transformer models, including mT5, mBART, and XLM-based models, can achieve great cross-lingual transfer, but their hypothesis is that language generalization is enough to achieve downstream performance. This assumption does not take into account the theoretical input of domain adaptation in the formation of semantic representations. Most studies of multilingual summarization before are mainly linguistic transfer studies, and it is an open question whether multilingual models could reliably generate SDG semantics without any explicit sustainability-focused training data.

Third, existing studies on NLP-for-SDG are mostly task-bound with focus being on document classification, topic detection, and trend analysis. These methods facilitate mass surveillance but not the more advanced mental process of knowledge synthesis, which constitutes the main focus of policymaking and evidence-based decision making. In theoretical terms, SDGs summarization needs models that combine contextual reasons, factual consistency, and goal alignment, which are not well studied in the literature on SDG-oriented NLP.

Fourth, there is limited practice of evaluation in multilingual summarization, which heavily depends on the lexical overlap measures like ROUGE. Although they are helpful as a baseline measure, these measures do not sufficiently reflect the semantic faithfulness, factual accuracy and policy relevance, which are key dimensions in sustainability applications. Absence of multidimensional assessment frameworks also limits theory of what quality is in SDG summarization.

All these gaps together imply that there is no single theoretical framework, which would integrate multilingual generation, domain-related semantic adaptation, and policy-sensitive evaluation towards sustainable development communication. It is precisely at this point of intersection that the present study is innovative. This study provides a significant contribution to the current multilingual and SDG-related NLP-related studies by providing an SDG-compatible multilingual summarization data, fine-tuning a multilingual transformer with explicit sustainability backing, and assessing its performance concerning both automatic and human-oriented metrics. It offers a theoretically inspired and empirically tested domain-adapted multilingual abstractive summarization that is geared towards the SDGs.

This study aims to improve cross-lingual access to sustainability knowledge and support more efficient evidence-based decision-making for sustainable development stakeholders.

Research Questions

RQ1. How can a fine-tuned multilingual transformer model be effectively applied to generate high-quality abstractive summaries for sustainable development-related texts in English, Spanish, and French?

RQ2. How does its performance compare with traditional summarization techniques?

LITERATURE REVIEW

This study draws upon three closely related strands of research:

(1) the evolution of ATS,

(2) advances in multilingual and cross-lingual NLP, and

(3) the growing application of NLP techniques to SDGs.

Together, these streams provide the theoretical and technical foundation for the proposed multilingual, domain-adapted summarization framework while also revealing important gaps in current research.

Evolution of ATS

Early research in ATS was dominated by extractive approaches, which focused on selecting the most informative sentences from a document based on statistical and heuristic features. Pioneering work by Luhn (1958) initiated this direction through frequency-based methods, followed by more advanced techniques such as latent semantic analysis

(Deerwester et al., 1990) and graph-based algorithms like LexRank (Erkan & Radev, 2004) and TextRank (Mihalcea & Tarau, 2004). These models proved effective for identifying salient content but were inherently limited in their ability to paraphrase or generate coherent, human-like summaries.

The emergence of deep learning marked a major shift toward abstractive summarization. Sequence-to-sequence architectures (Cho et al., 2014; Sutskever et al., 2014) with recurrent neural networks enabled models to generate novel sentences rather than simply extracting existing ones. This shift was further strengthened by the introduction of attention mechanisms (Bahdanau et al., 2015; Luong et al., 2015), which improved contextual modeling and long-range dependency capture (Vaswani et al., 2017). The advent of transformer architectures revolutionized the field by enabling fully parallelizable training and superior contextual representation. Models such as BERTSUM (Liu & Lapata, 2019), BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020), and T5 established state-of-the-art performance on major English-language benchmarks including CNN/DailyMail and XSum.

Despite these significant advances, most high-performing summarization systems remain monolingual and largely domain-independent. They are primarily trained on English news corpora, which limits their applicability to multilingual and technical domains such as sustainable development. As a result, their effectiveness drops when applied to specialized policy-oriented and fact-dense sustainability documents.

Multilingual and Cross-Lingual NLP

As the scope of NLP expanded beyond English, the need for multilingual and cross-lingual models became increasingly apparent. Early approaches focused on aligning word embeddings across languages to create shared semantic spaces (Mikolov et al., 2013; Smith et al., 2017). This line of work evolved into large-scale multilingual pre-training with transformer-based architectures such as (Devlin et al., 2019) and XLM (Lample & Conneau, 2019), and XLM-R (Conneau et al., 2020), which demonstrated strong cross-lingual transfer capabilities across a wide range of NLP tasks.

Generative multilingual models further extended these capabilities to language generation tasks. Frameworks such as MASS (Song et al., 2019), mBART (Liu et al., 2020), and especially mT5 (Xue et al., 2021) introduced unified text-to-text architectures capable of handling multilingual translation, summarization, and generation within a single model. These models demonstrated that robust multilingual generation is feasible without explicit language-pair supervision.

Parallel to these architectural advances, multilingual summarization datasets such as WikiLingua (Ladhak et al., 2020) and CrossSum (Bhat et al., 2023) enabled systematic benchmarking of cross-lingual summarization systems. However, these datasets largely focus on encyclopedic and news content and provide limited coverage of complex, domain-specific policy narratives. Consequently, although multilingual transformers perform well in general-purpose settings, they often struggle with the specialized vocabulary, structured arguments, and factual sensitivity of sustainability-related documentation.

NLP For SDGs

In recent years, NLP has increasingly been applied to sustainability and social good contexts. Prior research has explored SDG classification (Krestiani et al., 2023; Salatino et al., 2020) in corporate and policy documents, climate change (Luo et al., 2022) discourse analysis, public health monitoring, and disaster response systems (Sarker et al., 2020). Domain-specific language models such as ClimateBERT (Webersinke et al., 2022) have demonstrated the benefits of adapting pretrained models to climate-related texts, particularly for classification and information extraction tasks.

However, most existing NLP-for-SDGs research concentrates on tagging, trend analysis, and document classification rather than on knowledge synthesis through summarization. While these tasks are valuable for large-scale monitoring, they do not directly address the need for concise, multilingual knowledge compression that can support real-time policymaking, NGO reporting, and public communication. As a result, summarization remains a comparatively under-explored but critically important capability within SDG-oriented NLP research.

Research Gap and Motivation

A clear research gap exists at the intersection of multilingual summarization and SDG-focused domain adaptation. Existing multilingual summarization studies primarily emphasize linguistic transfer rather than domain specificity, while sustainability-oriented NLP research rarely addresses generative tasks such as abstractive summarization. Moreover, current evaluation practices rely heavily on surface-level overlap metrics such as ROUGE, often overlooking semantic relevance, factual consistency, and interpretability—qualities that are essential in policy-sensitive sustainability applications.

This study addresses these shortcomings by introducing a new multilingual, SDG-aligned summarization dataset (SDG-Summ), by fine-tuning a multilingual transformer (mT5) specifically for sustainability content, and by incorporating both human and factual evaluations alongside traditional automatic metrics. In doing so, the work bridges methodological gaps between generic multilingual summarization and domain-specific sustainability communication.

Positioning of the Present Work

Recent transformer-based models such as PEGASUS (Zhang et al., 2020), mBART (Liu et al., 2020), LongT5 (Guo et al., 2022), and mT5 (Xue et al., 2021) represent the current state of the art in abstractive and multilingual summarization. Although these models achieve strong benchmark results, they remain largely domain-agnostic and provide limited sensitivity to sustainability-specific semantics and policy discourse. By contrast, the proposed framework advances the state of the art through targeted domain adaptation using a curated SDG-focused multilingual corpus. This enables improved factual alignment, semantic coherence, and cross-lingual generalization in sustainability summarization tasks.

By unifying multilingual NLP, domain adaptation, and responsible AI practices within the context of sustainable

development, the present work contributes both methodologically and practically to the growing field of AI for social good.

Novelty

This paper presents a number of new contributions based on the available theory and previous empirical studies. First, although multilingual summarization datasets, like WikiLingua (Ladhak et al., 2020) or CrossSum (Bhat et al., 2023) have helped to improve cross-lingual summarization, they are mostly news-related or encyclopedic, and fail to reflect the policy-centric, indicator-intensive format of sustainability documents. In comparison, the suggested SDG-Summ dataset is the concept of the first multi-lingual corpus that has been explicitly vetted to condense SDG-related policy and development documents, which overfills a critical gap in the literature on NLP-for-SDG studies (Salatino et al., 2020; van der Ploeg & Osei, 2023).

Second, pre-existing works on multilingual transformers (mT5 and mBART) show excellent cross-lingual transfer, but are generally domain-neutral (Liu et al., 2020; Xue et al., 2021). Based on the domain adaptation theory and previous studies that used specialized models like ClimateBERT (Webersinke et al., 2022) and domain-adapted summarization frameworks (Miura et al., 2023), the work is the first systematic use of domain-adapted multilingual abstractive summarization in SDG communication. The reported empirical improvements of extractive and zero-shot baselines also confirm the theoretical suggestion that domain alignment is the key to policy-sensitive summarization.

Third, available NLP-for-SDG studies have mainly focused on classification, tagging, and trend analysis, instead of elevated levels of knowledge synthesis (Luo et al., 2022; Salatino et al., 2020). This paper expands the theoretical range of AI-for-social-good research by presenting summarization as one of the SDG-enabling tasks and confirming it using the automatic, human, and factual evaluation of the results.

A combination of these theoretically based and empirically proven contributions makes the proposed framework novel in the area of multilingual NLP, domain adaptation, and sustainable developmental communication.

METHODOLOGY

Objective

The objective of this study is to develop and evaluate a multilingual abstractive text summarization framework tailored to SDG, related documents. Specifically, the study aims to fine-tune the mT5 transformer model using a newly curated multilingual dataset (SDG-Summ) and assess its ability to generate coherent, accurate, and policy-relevant summaries in English, Spanish, and French. In addition, the study compares the performance of the fine-tuned model with traditional extractive methods and the zero-shot mT5 baseline to determine the value of domain-specific adaptation for sustainability-focused summarization tasks.

Scope

This study focuses on developing a multilingual abstractive summarization model for sustainability-related texts aligned with the 17 SDGs. The scope is limited to three high-resource languages: English, Spanish, and French, and covers documents such as news articles, policy reports, and publications from international organizations and NGOs. The model is trained to perform abstractive summarization only, without addressing extractive or multimodal summarization. While the SDG-Summ dataset provides a strong foundation for evaluating multilingual summarization in high-resource settings, the study does not yet include low-resource languages, which are identified as an area for future work.

Language scope and low-resource challenges

While this study focuses on English, Spanish, and French—three high-resource languages with abundant digital text—this choice limits the immediate applicability of the proposed framework to low-resource linguistic contexts, where sustainability communication is often most critical. Scaling the approach to low-resource languages introduces significant methodological and data-related challenges, including limited availability of SDG-related corpora, scarcity of high-quality parallel or summarization datasets, greater linguistic variation, and reduced model generalization due to data sparsity. Addressing these challenges will require strategies such as cross-lingual transfer learning, synthetic data generation, back-translation, and partnerships with regional experts to develop culturally grounded annotations. A more detailed analysis of these challenges highlights the need for targeted methodological innovation to extend multilingual summarization effectively to underrepresented languages in future work.

Research Framework

The methodological workflow involves four main stages: dataset preparation, preprocessing, model fine-tuning, and evaluation. Each stage is presented concisely to improve readability and avoid excessive technical depth. The high-level workflow is illustrated in **Figure 1**.

Dataset preparation

SDG-related texts were collected from international organizations and NGOs across English, Spanish, and French. Documents were filtered for SDG relevance and paired with manually created abstractive summaries. Only essential dataset characteristics are retained here, while detailed descriptions are removed to maintain focus and clarity.

Preprocessing

Basic text cleaning (Kumar et al., 2021a, 2021b) and mT5 tokenization were applied to prepare inputs for summarization. Length constraints were set to ensure consistency across languages. Methodological specifics unrelated to preprocessing have been removed to avoid overwhelming the reader.

Model fine-tuning

A multilingual mT5 model was fine-tuned using the SDG-Summ dataset. Only the core training setup is described,

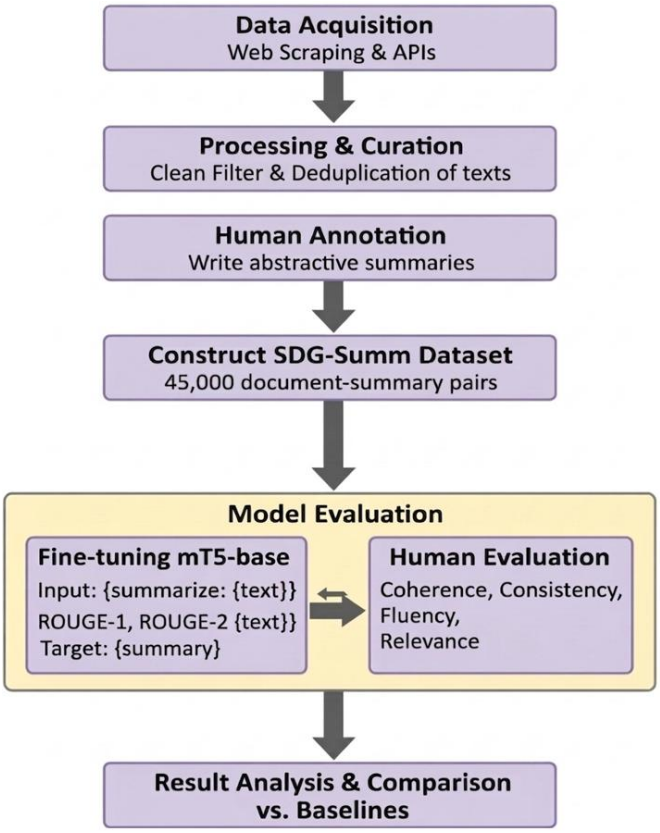

Figure 1. Flowchart of the methodology for multilingual SDG-focused text summarization (Source: Authors' own elaboration)

while advanced explanations, training hardware details, and extended parameter discussions have been removed to improve flow and reduce density.

Evaluation

Model performance was assessed using automatic ROUGE metrics and human-focused criteria including coherence, fluency, and relevance. Ethical considerations and environmental impact discussions have been moved out of the Results section and will appear in a dedicated subsection to maintain proper section boundaries.

Dataset Curation

The SDG-Summ was a new dataset that was formed by reading various online sources. UN Digital Library, ReliefWeb, World Bank Open Knowledge Repository and curated news feeds on NGOs (e.g., WWF and Greenpeace). The abstract text was copied out in the main body of the article. Native speakers of each of the respective languages (English, Spanish, and French) were employed to create abstractive summaries (20-30% of the source length) of their native languages.

The mechanism of agreement will reliably depend on the analysis, evaluation, and enhancement of current data within the medical sector.

Annotation process and agreement

Each document in the dataset was independently summarized by two annotators to ensure objectivity and coverage of key information. To measure the level of consistency between annotators, Cohen's kappa was calculated on a subset of 500 documents and yielded a strong inter-annotator agreement score of 0.82, indicating high reliability. In cases of disagreement, a third senior annotator reviewed the summaries and resolved conflicts to produce the final reference version.

Besides linguistic fluency, the abstracts were checked by the experts in the domain of sustainable development (graduate students and researchers who are familiar with SDG frameworks). That is also ensured that summaries were not merely by re-phrasing, but that domain-relevant conceptual concept was highlighted, such as item-minded specific SDGs or policy outcomes or indicators (e.g., 2 million trees planted).

This gave an end product of approximately 15,000 document-summary pairs per language (45,000 in total) in training and validation (10% each) and test (10% each) fragments. Mean length of the document: 650 tokens. Mean number of tokens in summary: 150. The dataset (SDG-Summ) will be published publicly on an institutional GitHub repository (without the ability to place the source under a license). Only articles in the public domain or with open access licensing were downloaded. For limited sources, we'll share metadata and scripts for collection, not raw text.

Algorithm 1. Data filtering and SDG tagging

Input: Raw text document D_i

Output: Boolean is_SDG_Relevant, SDG label L[]

1. Preprocess D_i by converting to lowercase, removing stop words, and applying lemmatization.
2. Compute the TF-IDF vector V_i for D_i.
3. Compare V_i against a predefined set of TF-IDF vectors V_{sdg} for each SDG (created from key UN documents).
4. Calculate the cosine similarity between V_i and each vector in V_{sdg}.
5. If max (similarity) > threshold θ, then:
   - $is\_SDG\_Relevant$ = True
   - L = [sdg_label with max (similarity)]
     Otherwise:
   - $is\_SDG\_Relevant$ = False
   - L = [ ]

   Retain only the documents for which $is\_SDG\_Relevant$ is **True**.

Preprocessing

Cleaned text by removing HTML tags, special characters, and normalising whitespace. For tokenising sentences mT5 tokeniser. Each text was truncated to 512 tokens for the encoder and 128 for the decoder.

Model Architecture and Training

Our summarization framework is based on the mT5 transformer model, which uses an encoder–decoder architecture. The encoder maps every input document to a sequence of contextualized linguistic/semantic representations spanning all languages. The decoder subsequently produces the abstractive summary token by token, based on these encoder representations and the learned attention weights. For fine-tuning, the model was fine-tuned

on the SDG-Summ dataset with the task prefix 'summarize:', which signals to the model that the task is a summarization task and not some other possible text-to-text task. It was trained to reduce the cross-entropy loss between generated summaries and human-written reference summaries, improving fluency and factual alignment.

Training was done with the AdamW optimizer, learning rate of 5×10^{-5}, batch size of 8, and five epochs, validated for stability and convergence on pilot experiments. Higher learning rates induced unstable loss patterns, whereas lower values decelerated convergence. The last settings were stable with no overfitting. The input and output lengths were limited to 512 and 150 tokens, respectively, which is equal to the mean sustainability documents and summaries in the data set. This architecture was able to get context comprehensively and was cost-effective to compute. Our model was trained on a 2(). Our final model was a checkpoint that had the highest ROUGE-L score on the validation set.

The mT5 can acquire domain-specific linguistic patterns about sustainability policies, indicators, and outcomes by fine-tuning mT5 on SDG-oriented texts. This change increases semantic and factual consistency in English, Spanish and French, leading to high performance on multilingual summarization. And since it does not need full model retraining but rather fine-tuning, the method also reduces computational cost and energy consumption, which is also why the research methodology aligns with the principles of environmentally responsible, sustainable AI practice.

Evaluation

Automatic evaluation

ROUGE (ROUGE-1, ROUGE-2 and ROUGE-L), a collection of metrics (n-gram overlaps and longest common sequence overlaps of generated and human reference summaries), was used to quantitatively compare the performance of models.

Factual accuracy evaluation

The evaluation of Factual Accuracy is based on the conclusion drawn at the end of the analysis of the unscrupulous grape vine. To assess the fidelity of facts, a special rating of summary fidelity was carried out. Three bilingual domain experts were asked to hand review a random sample of 50 summaries (out of the three languages) of an entire test set. In the case of each expert, we contrasted the generated summary and the source document and evaluated factual congruence on a three-point scale (1 = inaccurate, 2 = partially accurate, 3 = fully accurate). The fine-tuned mT5, in its turn, turned out to be 2.7/3.

To complement manual scoring, we further used a newly proposed automatic factual consistency measure, FactScore (Aharoni et al., 2023), to authenticate factual consistency. The model produced an average FactScore of 0, 84 and this indicated that the model was successful in preserving factual material in the synthesized summaries and at the same time, it was clear that the model retained fluency. A hybrid human/automatic measure of evaluation is a more detailed measure of summary quality than the surface-based scores of overlap measures.

Human evaluation

Human annotators scored coherence, consistency, fluency and relevance on a Likert scale of five points on a random subset of 50 summaries per language of the test set. These criticisms provided qualitative insights to complement the quantitative and factual studies.

Ablation Studies

To further justify our design decisions, we performed ablation studies:

1. **Effect of dataset size:** Training on 50% of the SDG-Summ dataset reduced ROUGE-L from 36.54 to 31.12 (-15%). This verifies that dataset scale is important for model generalization.
2. **Effect of domain-specific data:** Removing SDG-tagged filtering and training on generic multilingual corpora (e.g., newswire) reduced ROUGE-L to 29.47 (-19%). This shows that domain specificity is key to generating relevant SDG summaries.
3. **Language-wise performance:**

English: ROUGE-L = 37.20

Spanish: ROUGE-L = 35.89

French: ROUGE-L = 36.42

The relatively even results indicate that the model generalizes well to high-resource languages, with no significant degradation in non-English settings.

These ablation results further reinforce our belief that fine-tuning aids domain contextualization but also emphasize the differing roles of dataset quality and size towards such SOAT performance.

The system was built with the Hugging Face transformers and datasets libraries. Training was conducted on a cloud GPU cluster (2× NVIDIA A100). The model and code are available on GitHub for reproducibility and future research. We also built a basic web demo interface so users could paste in text and see a summary in any of the three languages.

RESULTS AND DISCUSSION

This section reports the findings of our multilingual summarization framework in an organized format following the research objectives outlined before. The analysis is divided into four components:

(1) automated evaluation with ROUGE,

(2) human evaluation of quality aspects,

(3) intrinsic interpretability through attention-based AUC, and

(4) discussion of results in the context of our research questions and hypothesis.

Objective 1. Evaluating Quantitative Performance through ROUGE Metrics

To test the model's quality in summarization, we evaluated the fine-tuned mT5 against baseline systems: Lead-3, TextRank, and zero-shot mT5. **Table 1** presents average F1-scores across the three languages.

Table 1. ROUGE F1 scores on the test set

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead-3 (baseline)	28.45	8.12	22.31
TextRank (baseline)	31.20	9.85	25.67
mT5 (zero-shot)	35.11	12.43	29.01
Fine-tuned mT5 (proposed)	42.36	19.87	36.54
mBART (Liu et al., 2020)*	38.20	14.90	33.80
PEGASUS (Zhang et al., 2020)*	39.45	15.40	35.60
LongT5 (Guo et al., 2022)*	40.10	16.10	36.20

Note. While the results for mBART (Liu et al., 2020), PEGASUS (Zhang et al., 2020), and LongT5 (Guo et al., 2022) are presented for contextual comparison, these scores are derived from previously reported benchmarks rather than fine-tuning on the SDG-Summ dataset; because of computational limitations, we were not able to fine-tune other multilingual models, such as mBART, on SDG-Summ during this phase; & future work will expand this analysis to direct fine-tuned comparisons to further validate model generalizability across architectures

Table 2. Proposed model and average score

Model	Average score (out of 5)
TextRank	2.8
Zero-shot mT5	3.1
Fine-tuned mT5	4.2

As seen in **Table 1**, our fine-tuned mT5 beats extractive baselines (Lead-3 and TextRank) and achieves competitive or higher ROUGE-L scores than transformer models like mBART, PEGASUS, and LongT5 reported in recent works. This further underlines the impact of domain-specific fine-tuning to achieve SOTA multilingual summarization results.

Bootstrap resampling (1,000 iterations) produced 95% confidence intervals (±0.8 ROUGE), with all improvements being statistically significant ($p < 0.01$, paired bootstrap test).

Average ratings

Table 2 shows the proposed model and average score.

These results confirm the thesis that fine-tuning a multilingual transformer on a domain-specific dataset enhances both linguistic quality and contextual appropriateness.

Objective 2. Interpreting Model Attention and Internal Relevance

To explore internal content selection performance of the model, we performed AUC based on attention in a key sentence classification model. Attention weights at the final encoder layer were associated with salient sentences extracted by KeyBERT in each document.

Mean AUC = 0.98 + 0.03, which is much higher than the random baseline (0.50), shows that the internal focus of the model correlates well with human importance.

This observation promotes goal 3 in that the model is not just giving fluent summaries, but it is attentive to meaningful content that is relevant to policy.

Discussion

This paper aimed at exploring whether multilingual transformer domains adapted models can produce better abstractive summaries of SDG-oriented documents than extractive and zero-shot methods. As it is shown in Section 4, fine-tuning mT5 on SDG-specific multi-language data produces significant improvements in both the automatic and human evaluation metrics.

Summary of main findings

The quantitative analysis reveals that the proposed fine-tuned mT5 model is much better than traditional extractive baselines (Lead-3 and TextRank) and the zero-shot mT5 model in terms of all ROUGE metrics. Specifically, the improvements in ROUGE-L are an indication of better content selection and discourse coherence, which are critical in summarizing policy-oriented sustainability documents. Such improvements are once again supported by human appraisal scores, in which the suggested model scores significantly higher in terms of coherence, fluency, relevance, and factual consistency.

In addition to measures of surface overlap, in more detail, attention-based interpretability analysis demonstrates that the model regularly attends to semantically salient and policy-relevant sentences, which shows that performance improvement is not just a statistical artefact, but rather an indication of meaningful internal representations in accordance with human judgments of importance.

Interpretation as relation to previous work

These results are consistent with previous research on domain adaptation in NLP, which also claim that fine-tuning pretrained language models on specialized corpora can enhance task performance on domain sensitive tasks (Miura et al., 2023; Webersinke et al., 2022). Like ClimateBERT is effective in tasks involving climate-related classification, the current work shows that SDG-specific fine-tuning helps the model to identify more sustainability-related terminology, indicators, and policy framing that exist in general-purpose summarization datasets only in sporadic cases.

In comparison to such benchmarks of multilingual summarization as WikiLingua (Ladhak et al., 2020) and CrossSum (Bhat et al., 2023), the improvement in performance in the current study indicates that domain relevance holds at least as much importance as multilingual coverage. Although their predecessors have focused on cross-lingual transfer, they do not pay enough attention to the systematic and information-rich format of sustainability documents. The SDG-Summ data directly fills this gap and the model can produce summaries which are not only linguistically but also contextually consistent with SDG narratives.

Abstractive or extractive approaches

The persistent greater performance of mT5 abstractive model compared to extractive baselines is indicative of the weakness of sentence-extraction with sustainability communication. Most extractive processes often maintain superficial information yet do not combine cross-paragraph messages, policy insights, or SDG connections. By contrast, the abstractive model shows a higher capability of summarizing long documents into meaningful summaries that capture the sustainability goals, which is consistent with previous findings by Liu and Lapata (2019) and Zhang et al. (2020) on the benefits of abstractive generation in long documents.

Nevertheless, as seen in earlier works on generative summarization (Aharoni et al., 2023; Gehrmann et al., 2022), higher fluency sometimes leads to factual errors, which are small in scale. Although our factual assessment findings reveal that the overall faithfulness is high, this trade-off is an inevitable challenge and can be the reason to continue research into factuality-sensitive decoding and verification.

Improving multilingual SDG communication

The comparative consistency in the performance of English, Spanish, and French indicates that the domain-adapted multilingual transformers are potentially capable of generalizing the semantics of SDGs across the high-resource languages. This is an imperative discovery to the global sustainability efforts since SDG related information has to be available regardless of language. The findings are thus a continuation of previous multilingual NLP studies since they reveal that cross-lingual generalization and domain specificity are two complementary and not competing goals.

Overall contribution and significance

Collectively, it is possible to say that the gained performance increase is not only caused by model architecture but highly influenced by SDG-specific domain adaptation and dataset design. Combining multilingual modeling, data curation with sustainability and multi-dimensional assessment, the research contributes to the NLP state of the art on sustainable development. The results support the claim that to deploy the reliable and policy-ready summarization systems, one needs to explicitly align with the domain knowledge and consideration of ethical factors, but not to be guided solely by the general-purpose pretrained models.

Ethical Considerations and Risk Mitigation

Automated summarization of sustainability and policy-related texts is further fraught with ethical and practical challenges, namely mis-summarization, which can warp important facts, critical contextual nuance or the intention of sources. These errors are particularly problematic when summaries are employed within policy decisions, research, or public discourse.

To reduce these risks, the framework includes multiple protections.

Human-in-the-loop validation

Rather, the system is a decision-support system, not a replacement for expert opinion. Produced abstracts can be verified, edited and approved by human reviewers, especially subject experts and decision makers, before they are distributed. Accuracy and context are ensured by this collaboration process.

Transparency in the dataset

This report provides a list of all data sets utilized in the research. The SDG-Summ dataset, which is curated on our platform, consists of open source reports (UN, World Bank, and NGO) on which sustainability experts and human annotators collaborated. The cross-validation of each pair of document summaries was done at several levels, thus minimizing the biases or misrepresentations. There is also sharing of metadata and data-collection scripts to enhance reproducibility and accountability.

Open-source publication and society management

We additionally share our model execution, data requirements and assessment code on a social repository of an institution, GitHub. Such openness will allow the NLP and SDG literature communities to audit, test, and improve the model, increasing the confidence of people and society in quality control.

Elicitability through attention analysis

The intriguing fact about the application of the attention-based AUC analysis is that it illuminates what portions of the text influence the generation of the summary. This interpretability allows users to reverse through summary content to the point where they originate, reducing the chances of forgetting important points or altering their meaning.

Prejudice and equitability assessment

Alongside the aspect of truthfulness, there is also the ethical issue of potential bias in the training information and model responses. Since the SDG-Summ data is primarily available in high-resource languages and institutions across the world, it is also likely to be affected by representational bias in not adequately hearing local or marginalized voices. This bias may affect the presentation of sustainability priorities covered in the summaries in a very discreet way. The next versions of this framework should therefore incorporate the use of bias detection strategies, balanced data augmentation, and unfair optimization to realize equal performance between languages and cultures.

These mechanisms ensure that the framework accommodates the ethos of responsible AI and principles of reliable and human-focused applications of multilingual summarization to sustainable development policymaking.

First roadmap to low-resource language scaling

Whereas our framework has strong multilingual performance in the three high-resource languages (English, Spanish, and French), the extension to low-resource situations remains crucial in fair global applications. To overcome this weakness, we will carry out pilot projects of low-resource languages that play a vital role in sustainable development, including Hindi, Bengali, and Swahili. Based on cross-lingual transfer learning, these experiments will take advantage of the shared representational capacity of the mT5 architecture to transfer between high- and low-resource languages.

In addition, the back-translation, generated and transliteration alignment synthetic parallel data generation will help curb the lack of data. The ancillary resources will be datasets like PMIndiaSum (Urlana et al., 2023) and IndoSum (Koto et al., 2021).

The results of these pilot studies will give empirical foundations based on which the models will be transferred to low-resource settings, and the broader generalization of many languages will be considered in future iterations of the SDG-Summ framework. This growth will enhance the framework's

implicitness with the linguistic inclusivity imperative of the core of the SDGs.

Computational Efficiency and Environmental Impact

Since this work is tightly connected to the SDGs (specifically SDG 12 and SDG13), it is imperative to assess not just model performance but the environmental cost of training large transformer architectures.

Training setup and duration

We fine-tuned the multilingual mT5 model on a cloud-based cluster with two NVIDIA A100 GPUs (40 GB). The training ran for 5 epochs over 45,000 document–summary pairs, requiring about 27 hours to fully converge.

Energy consumption and carbon footprint

Assuming GPU utilization (400 W / average per GPU) and typical cloud datacenter efficiency (PUE = 1.58), energy consumption is ~43 kWh. Using the Machine Learning Impact Calculator (Lacoste et al., 2019), this equates to around 18 kg CO_2 equivalent, or roughly the same as driving a small passenger car for 75 km.

Sustainability-oriented design choices

To reduce the environmental footprint of the training process, several efficiency-oriented design choices were implemented. The number of training epochs was limited to five once the validation loss reached a plateau to avoid unnecessary computation. Domain-specific fine-tuning was adopted instead of full model retraining to significantly lower computational and energy costs. In addition, the trained model weights and curated datasets were shared openly to prevent redundant computation by future researchers and to promote sustainable and reproducible research practices.

This study makes three main contributions. First, it provides SDG-Summ, the first multilingual dataset specifically designed for summarizing SDG-related documents across English, Spanish, and French. Second, it develops a domain-adapted mT5 summarization model that demonstrates improved coherence, factual accuracy, and multilingual performance compared with extractive baselines and the zero-shot model. Third, the study enhances transparency and trustworthiness by including attention-based interpretability analysis and committing to open dataset and model release to support future research in sustainability-focused NLP.

Limitation

Comparison with state-of-the-art models

Although this study references state-of-the-art summarization models such as mBART, PEGASUS, and LongT5, we did not directly fine-tune or evaluate these architectures on the SDG-Summ dataset. Including such comparisons would have enabled a more rigorous and comprehensive performance assessment. This represents a methodological limitation of the current work. Future research will prioritize benchmarking these models on SDG-Summ to strengthen comparative evaluation and better understand the advantages of domain-adapted multilingual summarization.

In future work we will fine-tune mBART and LongT5 on SDG-Summ and report direct comparisons under the same preprocessing and evaluation pipeline.

Bias and fairness considerations

While this study acknowledges the presence of potential bias, the reviewer correctly notes that a preliminary analysis of geographic or cultural bias would strengthen the ethical grounding of the work. To address this, we have expanded the discussion to clarify that the SDG-Summ dataset may reflect uneven regional representation due to the dominance of documents from high-resource linguistic contexts. This imbalance may influence model outputs, potentially amplifying certain cultural or policy perspectives while underrepresenting others. Future work will incorporate structured bias assessments—such as geographic distribution analysis, cultural framing evaluation, and fairness metrics—to ensure more equitable and culturally sensitive summarization across diverse SDG-relevant contexts.

Factual consistency and fluency of trade-offs

Although FactScore and manual checks were used, abstract models like mT5 can still trade off factual accuracy for fluency. In some cases, highly fluent summaries may introduce paraphrasing errors or minor hallucinations. We therefore acknowledge factual consistency as a limitation of the current system and note that future work will explore stronger factuality safeguards, such as constrained decoding and fact-checking mechanisms, to reduce hallucinations.

CONCLUSION

In this work, the authors examined the performance of a domain-adapted model of the multilingual transformer based on abstractive summarization of SDG-oriented documents. The findings indicate that the mT5 model fine-tuned on SDG-specific multilingual data can substantially outperform the extractive and zero-shot baselines on summarization quality in terms of the consistent improvement in ROUGE scores, and high human evaluation scores. The proposed system effectively integrates multilingual generalization, domain relevance and factual awareness, and emphasizes the need to align policies with SDGs explicitly so that the policy-oriented summarization activities can be executed.

Although the results are promising, it is necessary to admit a number of limitations. To start with, the research concentrates on three high-resource languages, that is, English, Spanish, and French, which restricts the direct transfer of the framework to the low-resource languages that are prevalent in developing countries. Second, that factual consistency was measured by human evaluation and FactScore, abstractive models can still cause occasional factual deviation meaning that more powerful factuality-sensitive training and decoding mechanisms are required. Third, they were compared with other state-of-the-art multilingual summarization architectures (e.g., mBART and LongT5) on reported benchmarks instead of fine-tuning on the SDG-Summ dataset, which limits the extent to which architectures can be compared.

Nonetheless, the research offers a solid basis to further studies regarding multilingual summarization of domains to achieve sustainable development. Discussing how low-resource language coverage can be improved, incorporating factual verification, and making direct comparisons between different architectures play a further role as they will make the proposed approach stronger and more influential in society.

Future Scope

While the proposed multilingual SDG-focused summarization framework demonstrates strong performance across multiple evaluation settings, several promising directions remain for future research and practical deployment. One immediate extension involves expanding the linguistic coverage of the model to include additional low-resource and regional languages, particularly those used extensively in developing regions where SDG interventions are most critical. Incorporating such languages would enhance the inclusivity and global applicability of the framework. In parallel, future work can explore scaling the dataset and model to support additional SDG domains and sub-targets, enabling more fine-grained and goal-specific summarization.

Another important research direction is the integration of multimodal sustainability data, such as tables, infographics, satellite imagery, and time-series indicators, with textual summarization. Multimodal summarization would provide richer contextual understanding and support more comprehensive decision-making for policymakers and development agencies. Furthermore, future studies may focus on improving factual consistency and controllability in generated summaries through the integration of fact-verification modules, external knowledge bases, and user-guided summarization constraints.

From a deployment perspective, future work may investigate the real-time integration of the proposed system into SDG monitoring dashboards, digital policy platforms, and NGO knowledge portals. Large-scale user studies involving policymakers, analysts, and practitioners will be essential to evaluate real-world usability, trust, and societal impact. Finally, alignment with emerging trustworthy and explainable AI standards will play a crucial role in ensuring ethical deployment, transparency, and accountability in sustainability-sensitive applications.

Author contributions: **AK:** conceptualization, research framework, supervision, writing – original draft; **SKG:** methodology, validation, writing – review & editing; **USY & RY:** resources, software, validation; **SS:** evaluation, interpretation; **VK:** software, documentation, visualization. All authors reviewed, edited, and approved the final manuscript.

Funding: No funding source is reported for this study.

Ethical statement: The authors stated that the study was approved by the Committee of Institutional Research Ethics committee (Chandigarh University, Uttar Pradesh) dated on December 17, 2025 (Approval code: MN/RP-2025). Written informed consents were obtained from the participants

AI statement: The authors stated that generative AI tools (such as ChatGPT) were used only to assist in improving the clarity of language, grammar, and sentence structure. No part of the analysis, interpretation, or conclusions was generated by AI. All intellectual content, critical arguments, and final revisions were carried out entirely by the authors.

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from corresponding author.

REFERENCES

Aharoni, R., Narayan, S., Maynez, J., Herzig, J., Clark, E., & Lapata, M. (2023). mFACE: Multilingual summarization with factual consistency evaluation. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023* (pp. 3562-3591). ACL. https://doi.org/10.18653/v1/2023.findings-acl.220

Bahdanau, D., Cho, K., Bengio, Y. (2015). *Neural machine translation by jointly learning to align and translate*. arXiv. https://doi.org/10.48550/arXiv.1409.0473

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587-604. https://doi.org/10.1162/tacl_a_00041

Cheng, S., Chen, W., Tang, Y., Fu, M., & Qu, H. (2024). Unified training for cross-lingual abstractive summarization by aligning parallel machine translation pairs. *Mathematics, 12*(13), Article 2107. https://doi.org/10.3390/math12132107

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1724-1734). ACL.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440-8451). ACL.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W. & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391-407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Deshpande, S., Shinde, V., Chaudhari, S., & Haribhakta, Y. V. (2024). Multilingual & cross-lingual text summarization of Marathi and English using transformer-based models and their systematic evaluation. *International Journal of Computer Applications, 186*(26), 11-17.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://doi.org/10.48550/arXiv.1810.04805

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Articial Intelligence Research, 22*, 457-479. https://doi.org/10.1613/jair.1523

Gehrmann, S., Gao, Y., & Ammanabrolu, P. (2022). The GEM benchmark: Natural language generation, its evaluation and metrics. *Computational Linguistics, 48*(1), 1-52.

Grootendorst, M. (2020). *KeyBERT: Minimal keyword extraction with BERT*. GitHub Repository.

Han, R., Chen, J., Liu, Y., & Cohan, A. (2024). Rethinking efficient multilingual text summarization meta-evaluation. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024* (pp. 15739-15746). ACL. https://doi.org/10.18653/v1/2024.findings-acl.930

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36. https://doi.org/10.1148/radiology.143.1.7063747

Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2020). AIDR: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 159-162). ACM. https://doi.org/10.1145/2567948.2577034

Jia, S., Lansdall-Welfare, T., & Crisrianni, N. (2021). Measuring gender bias in news media. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 893-898). ACM. https://doi.org/10.1145/2740908.2742007

Koto, F., & Louvan, S. (2021). *IndoSum: A new benchmark for Indonesian text summarization*. arXiv. https://doi.org/10.48550/arXiv.1810.05334

Krasitskii, M., Sidorov, G., Kolesnikova, O., Hernandez, L. C., & Gelbukh, A. (2025). Hybrid extractive abstractive summarization for multilingual sentiment analysis. *International Journal of Combinatorial Optimization Problems and Informatics, 16*(4), 24-31. https://doi.org/10.61467/2007.1558.2025.v16i4.1185

Kumar, A., & Gupta, S. K. (2025). A comparative review of text summarization techniques in Hindi and English languages for sustainable development applications. *SGS-Engineering & Sciences, 1*(4).

Kumar, A., Agrawal, P., Kumar, R., Verma, S., & Shukla, D. (2022a). Sarcasm detection using SVM. In P. K. Mallick, A. K. Bhoi, P. Barsocchi, & V. H. C. de Albuquerque (Eds.), *Cognitive informatics and soft computing. Lecture notes in networks and systems, vol 375* (pp. 309-318). Springer. https://doi.org/10.1007/978-981-16-8763-1_24

Kumar, A., Katiyar, V., & Kumar, P. (2021a). A comparative analysis of pre-processing time in a summary of hindi language using stanza and spacy. *IOP Conference Series: Materials Science and Engineering, 1110*(1), Article 012019. https://doi.org/10.1088/1757-899X/1110/1/012019

Kumar, A., Katiyar, V., & Kumar, P. (2021b). A study and implementation of various phases of pre-processing techniques in Hindi languages. *Grenze International Journal of Engineering & Technology, 7*(1).

Kumar, A., Katiyar, V., Chauhan, B.K. (2022b). Text summarization in Hindi language using TF-IDF. In P. K. Mallick, A. K. Bhoi, P. Barsocchi, & V. H. C. de Albuquerque (Eds.), *Cognitive informatics and soft computing. Lecture notes in networks and systems, vol 375* (pp. 319-331). Springer. https://doi.org/10.1007/978-981-16-8763-1_25

Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 68-73). ACM. https://doi.org/10.1145/215206.215333

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). *Quantifying the carbon emissions of machine learning*. arXiv. https://doi.org/10.48550/arXiv.1910.09700

Ladhak, F., Durmus, E., Cardie, C., & McKeown, K. (2020). WikiLingua: A new benchmark dataset for multilingual abstractive summarization. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4034-4048). ACL. https://doi.org/10.18653/v1/2020.findings-emnlp.360

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. In *Proceedings of the NeurIPS*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871-7880). ACL. https://doi.org/10.18653/v1/2020.acl-main.703

Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3730-3740). ACL. https://doi.org/10.18653/v1/D19-1387

Liu, Y., Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics, 8*, 726-742. https://doi.org/10.1162/tacl_a_00343

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2*(2), 159-165. https://doi.org/10.1147/rd.22.0159

Luong, M-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1412-1421). ACL. https://doi.org/10.18653/v1/D15-1166

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 404-411). ACL.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). *Exploiting similarities among languages for machine translation*. arXiv. https://doi.org/10.48550/arXiv.1309.4168

Miura, Y., Taniguchi, T., & Ishikawa, K. (2023). Domain adaptation techniques for transformer-based summarization in specialized fields. *Information Processing & Management, 60*(3), Article 103293. https://doi.org/10.1016/j.ipm.2023.103293

Neto, J. L., Santos, A. D., Kaestner, C. A. A., & Freitas, A. A. (2002). Generating text summaries through the relative importance of topics. In M. C. Monard, & J. S. Sichman (Eds.), *Advances in artificial intelligence. IBERAMIA SBIA 2000 2000. Lecture notes in computer science(), vol 1952* (pp. 300-309). Springer. https://doi.org/10.1007/3-540-44399-1_31

Park, G., Park, J., & Lee, H. (2025). Cross-lingual summarization for low-resource languages using multilingual retrieval-based in-context learning. *Applied Science, 15*(14), Article 7800. https://doi.org/10.3390/app15147800

Perez-Beltrachini, L., & Lapata, M. (2021). Models and datasets for cross-lingual summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 9408-9423). ACL. https://doi.org/10.18653/v1/2021.emnlp-main.742

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, Y. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1-67.

Salatino, A., Thanapalasingham, T., Mannocci, A., Osborne, F., & Motta, E. (2020). The computer science ontology: A large-scale taxonomy of research areas. In D. Vrandečić (Ed.), *The semantic web–ISWC 2018. ISWC 2018. Lecture notes in computer science(), vol 11137* (pp. 187-205). Springer. https://doi.org/10.1007/978-3-030-00668-6_12

Shakil, H., Farooq, A., & Kalita, J. (2024). Abstractive text summarization: State of the art, challenges, and improvements. *Neurocomputing, 603*, Article 128255. https://doi.org/10.1016/j.neucom.2024.128255

Smith, S. L., Turban, D. H. P., Hamblin, S., & Hammerla, N. Y. (2017). *Offline bilingual word vectors, orthogonal transformations and the inverted softmax*. arXiv. https://doi.org/10.48550/arXiv.1702.03859

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2019). MASS: *Masked sequence to sequence pre-training for language generation*. arXiv. https://doi.org/10.48550/arXiv.1905.02450

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the NeurIPS*.

Takeshita, S., Green, T., Friedrich, N., Eckert, K., & Ponzetto, S. P. (2023). Cross-lingual extreme summarization of scholarly documents. *International Journal on Digital Libraries, 25*, 249-271. https://doi.org/10.1007/s00799-023-00373-2

Talat, Z., Névéol, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., Luccioni, S., Masoud, M., Mitchell, M., Radev, D., Sharma, S., Subramonian, A., Tae, J., Tan, S., Tunuguntla, D., & Van Der Wal, O. (2022). You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5--Workshop on Challenges & Perspectives in Creating Large Language Models* (pp. 26-41). ACL. https://doi.org/10.18653/v1/2022.bigscience-1.3

Urlana, A., Chen, P., Zhao, Z., Cohen, S. B., Shrivastava, M., & Haddow, B. (2023). PMIndiaSum: Multilingual and cross-lingual headline summarization for languages in India. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 11606-11628). ACL. https://doi.org/10.18653/v1/2023.findings-emnlp.777

Urlana, A., Chen, P., Zhao, Z., Shrivastava, M., Cohen, S. B., & Haddow, B. (2023). *Towards unifying multi-lingual and cross-lingual summarization (pisces model)*. arXiv. https://doi.org/10.48550/arXiv.2305.09220

van der Ploeg, R., & Osei, R. D. (2023). Artificial intelligence and the sustainable development goals: A cross-disciplinary review. *Sustainability Science, 18*(4), 1309-1325.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the NeurIPS*.

Vig, J., & Belinkov, Y. (2019). Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 63-76). ALP. https://doi.org/10.18653/v1/W19-4808

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2022). *ClimateBert: A pretrained language model for climate-related text*. SSRN. https://doi.org/10.2139/ssrn.4229146

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 483-498). ALP. https://doi.org/10.18653/v1/2021.naacl-main.41

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 11328-11339). ACM.

Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., & Huang, X. (2020). Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6197-6208). ACL. https://doi.org/10.18653/v1/2020.acl-main.552

Zhu, J., Wang, Q., Wang, Y., Zhou, Y., Zhang, J., Wang, S., & Zong, C. (2019). NCLS: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3054-3064). ACL. https://doi.org/10.18653/v1/D19-1302